(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 January 2004 (22.01.2004)

PCT

(10) International Publication Number
WO 2004/008371 A1

(51) International Patent Classification[7]: G06F 19/00, G01N 33/68

(21) International Application Number: PCT/IB2002/002731

(22) International Filing Date: 10 July 2002 (10.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: INSTITUT SUISSE DE BIOINFORMATIQUE [CH/CH]; Centre Médical Universitaire, 1, rue Michel-Servet, CH-1211 Genève 4 (CH).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: APPEL, Ron [CH/CH]; 26, route des Carres, CH-1252 Meinier (CH). HERNANDEZ, Patricia [CH/CH]; 13, av. Louis-Yung, CH-1290 Versoix (CH). GRAS, Robin [FR/FR]; 1, rue Jules-Ferry, F-74100 Annemasse (FR).

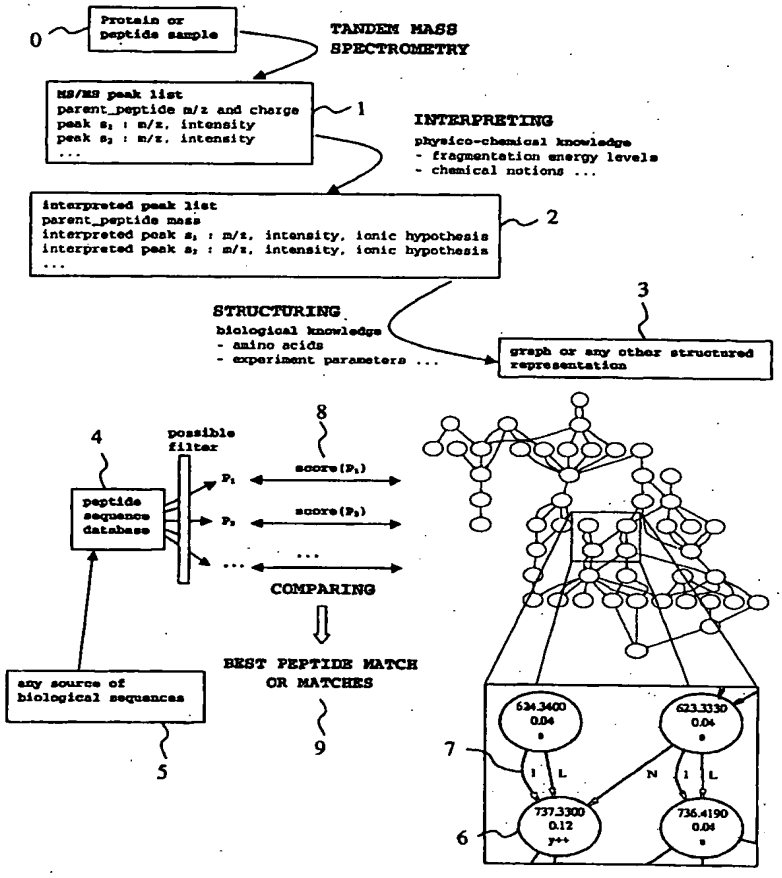
(74) Agent: DUCOR, Philippe; BMG Avocats, 8c, avenue de Champel, Case Postale 385, CH-1211 Geneva 12 (CH).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM); European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK,

[Continued on next page]

(54) Title: PEPTIDE AND PROTEIN IDENTIFICATION METHOD



(57) Abstract: Method for identifying peptides and proteins, starting from the corresponding tandem spectrometry data. More specifically, the method comprises performing tandem mass spectrometry on a sample containing one or more protein or peptide, reducing each resulting spectrum to a peak list, listing possible interpretations for said peak list into an interpreted peak list taking into account physico-chemical knowledge, structuring said interpreted peak list into a structured representation taking into account biological knowledge, matching said structured representation with a biological sequence database, and determining the best peptide match or matches within said database.

WO 2004/008371 A1

TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

PEPTIDE AND PROTEIN IDENTIFICATION METHOD

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to the field of proteomics and particularly to methods and systems for identifying peptides and proteins starting from tandem spectrometry data (MS/MS data) obtained experimentally. More specifically, the method comprises interpreting and structuring MS/MS data in a way allowing full exploitation of the information contained in it during matching of the structured data with biological sequence database.

The following references are either cited in the text or relevant to the prior art:

- Bafna V. and Edwards N. (2001). SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics Suppl 1, 13-21.

- Bairoch,A. and Apweiler,R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45-48.

- Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C., Yeh,L.S., Ledley,R.S., Janda,J.F., Pfeiffer,F., Mewes,H.W., Tsugita,A., and Wu,C. (2000). The protein information resource (PIR). Nucleic Acids Res. *28*, 41-44.

- Bartels C. (1990). Fast algorithm for peptide sequencing by mass spectrometry. Biomed. Environ. Mass. Spectrom. 19, 363-368.

- Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., and Wheeler,D.L. (2002). GenBank. Nucleic Acids Res. 30, 17-20.

- Bonabeau E., Dorigo M., and Theraulaz G. (1999). Swarm Intelligence. From Natural to Artificial Systems. Oxford University Press).

- Chen,T., Kao,M.Y., Tepel,M., Rush,J., and Church,G.M. (2001). A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. *8*, 325-337.

- Clauser K.R., Hall S.C., Smith D.M., Webb J.W., Andrews L.E., Tran H.M., Epstein L.B., and Burlingame A.L. (1995). Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. Proc Natl Acad Sci USA *92(11)*, 5072-5076.

- Dancik,V., Addona,T.A., Clauser,K.R., Vath,J.E., and Pevzner,P.A. (1999). De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol. 6, 327-342.

- Dorigo,M. and Di Caro,G. (1999). The Ant Colony Optimization Meta-Heuristic. In New Ideas in Optimization, D.M.G.F.E.Corne D., ed.

- Edman,P. (1970). Sequence determination. Mol. Biol. Biochem. Biophys. 8, 211-255.

- Eng J.K., McCormack,A.L., and Yates,I.J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid

sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976-989.

- Fenyo,D., Qin,J., and Chait,B.T. (1998). Protein identification using mass spectrometric information. Electrophoresis *19*, 998-1005.

- Fernandez-de-Cossio,J., Gonzalez,J., and Besada,V. (1995). A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. Comput. Appl. Biosci. *11*, 427-434.

- Fernandez-de-Cossio,J., Gonzalez,J., Betancourt,L., Besada,V., Padron,G., Shimonishi,Y., and Takao,T. (1998). Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by 'SeqMS', a software aid for de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *12*, 1867-1878.

- Fernandez-de-Cossio,J., Gonzalez,J., Satomi,Y., Shima,T., Okumura,N., Besada,V., Betancourt,L., Padron,G., Shimonishi,Y., and Takao,T. (2000). Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. Electrophoresis *21*, 1694-1699.

- Gatlin,C.L., Eng,J.K., Cross,S.T., Detter,J.C., and Yates,J.R., III (2000). Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. Anal. Chem. *72*, 757-763.

Gonnet G.H. A tutorial Introduction to Computational Biochemistry Using Darwin. 1992. E.T.H. Zurich, Switzerland.
Ref Type: Report

- Gras,R., Muller,M., Gasteiger,E., Gay,S., Binz,P.A., Bienvenut,W., Hoogland,C., Sanchez,J.C., Bairoch,A., Hochstrasser,D.F., and Appel,R.D. (1999). Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20, 3535-3550.

- Gras R., Gasteiger E., Chopard B., Müller M., and Appel R.D. New learning method to improving protein identification from peptide mass fingerprinting. 2000. 4th Siena 2D electrophoresis meeting.
Ref Type: Conference Proceeding

- Gras R. and Muller M. (2001). Computational aspects of protein identification by mass spectrometry. Current Opinion in Molecular Therapeutics *3*, 526-532.

- Hines W.M., Falick A.M., Burlingame A.L., and Gibson B.W. (1992). Pattern-based algorithm for peptide sequencing from tandem mass spectra of peptides. J. American Society for Mass Spectrometry 3, 326-336.

- Ishikawa,K. and Niwa,Y. (1986). Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. Biomed. Environ. Mass Spectrom *13*, 373-380.

- Johnson,R.S. and Biemann,K. (1989). Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. Biomed. Environ. Mass Spectrom *18*, 945-957.

- Johnson,R.S. and Taylor,J.A. (2000). Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. Methods Mol. Biol. 146, 41-61.

- Kennedy J. and Eberhart R.C. (2001). Swarm Intelligence. Morgan Kaufmann).

- Mann,M., Hojrup,P., and Roepstorff,P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom *22*, 338-345.

- Mann,M. and Wilm,M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. *66*, 4390-4399.

- Pappin D.D.J., Hojrup P., and Bleasby A.J. (1993). Rapid identification of proteins by peptide-mass finger printing. Curr Biol *3*, 327-332.

- Perkins D.N., Pappin D.D.J., Creasy D.M., and Cottrell J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.

- Pevzner,P.A., Dancik,V., and Tang,C.L. (2000). Mutation-tolerant protein identification by mass spectrometry. J. Comput. Biol. 7, 777-787.

- Pevzner,P.A., Mulyukov,Z., Dancik,V., and Tang,C.L. (2001). Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res. *11*, 290-299.

- Sakurai T., Matsuo T., Matsuda H., and Katakuse I. (1984). Paas 3: A computer program to determine probable sequence of peptides from mass spectrometric data. Biomed. Mass Spectrom. *11(8)*, 396-399.

- Siegel,M.M. and Bauman,N. (1988). An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. Biomed. Environ. Mass Spectrom. *15*, 333-343.

- Stoesser,G., Baker,W., van den,B.A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., Redaschi,N., Stoehr,P., Tuli,M.A., Tzouvara,K., and Vaughan,R. (2002). The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 30, 21-26.

- Tateno,Y., Imanishi,T., Miyazaki,S., Fukami-Kobayashi,K., Saitou,N., Sugawara,H., and Gojobori,T. (2002). DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res. *30*, 27-30.

- Taylor,J.A. and Johnson,R.S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. *11*, 1067-1075.

- Taylor,J.A. and Johnson,R.S. (2001). Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. Anal. Chem. *73*, 2594-2604.

- Wilkins M.R., Gasteiger E., Bairoch A., Sanchez J.C., Williams K.L., Appel R.D., and Hochstrasser D.F. (1999a). Protein identification and analysis tools in ExPASy server. Methods Mol Biol *112*, 531-552.

- Wilkins M.R., Gasteiger E., Wheeler C.H., Lindskog I., Sanchez J.C., Bairoch A., Appel R.D., Dunn M.J., and Hochstrasser D.F. (1999b). Multiple parameter cross-species protein identification using Multident - a world-wide web accessible tool. Electrophoresis *19*, 3199-3206.

- Yates,I.J.R., Eng J.K., and McCormak A.L. (1995). Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. *67(18)*, 3202-3210.

- YatesIII J.R., Eng J.K., Clauser K., and Burlingame A.L. (1996). Search of Sequence Databases with Uninterpreted High-Energy Collision-Induced Dissociation Spectra of Peptides. J. American Society for Mass Spectrometry *7*, 1089-1098.

- Zhang,W. and Chait,B.T. (2000). ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. Anal. Chem. *72*, 2482-2489.

2. Description of the Prior Art

Proteomics is the study of the proteins resulting from the expression of the genes contained in genomes. Due to important variations of protein expression between cells having the same genome, there are many proteomes for each corresponding genome. As a result, huge amounts of

information are involved, and the study of proteome is even more complex than the study of the genome.

A typical goal of proteomics is to identify the protein expression in a given tissue or cell under given conditions. An additional goal of proteomics is to compare the protein expression in the same tissue, cell or physiological fluid under varying conditions (for example disease vs control), and identify the proteins that are differently expressed.

In recent years, proteomics research has gained importance due to increasingly powerful techniques in protein purification/separation, mass spectrometry and identification techniques, as well as the development of extensive protein and nucleic databases from various organisms.

A traditional method for analyzing proteomes involves separation by 1-D and 2-D polyacrylamide-gel electrophoresis. The 1-D gel method is generally used to achieve a crude separation of cell lysates where the most abundant proteins can be separated and detected. 2-D gel electrophoresis is a more powerful method capable of separating out hundreds of protein spots, where the spot pattern is characteristic of protein expression. Typical separation criteria by gel electrophoresis include electrical charge (isoelectric point - pI) and molecular weight. Gel electrophoresis methods (1-D and 2-D) have nevertheless certain fundamental limitations for screening and identification of proteins. Notably, gel electrophoresis separations are slow and have a

limited resolution (i.e. can only distinguish between a limited number of proteins (spots)). In recent years, automation has allowed to manage larger quantities of data resulting from 2-D gel electrophoresis, as exemplified by US Pat. No. 5,993,627, US Pat. No.6,277,259, and WO 00/55636.

Higher resolution can be attained by other chromatography separation methods such as capillary electrophoresis, gas chromatography, micro-channel networks, liquid chromatography and high-pressure liquid chromatography (HPLC), used in complement to gel electrophoresis or alone. These methods allow the separation of greater numbers of proteins, even in hard conditions (low sample quantities, small molecular weight, highly basic or hydrophobic proteins...). Separation criteria include electrical charge and molecular weight as in gel electrophoresis, as well as hydrophobicity and other physico-chemical criteria.

After separation, the proteins must be identified, by sequencing or other means. Determining the sequence of amino acid residues in a protein was traditionally accomplished by means of N-terminal Edman degradation (Edman, 1970). Edman sequencing unfortunately requires important quantities of a protein (in the order of 10-100 pmols), which exceed the quantities obtained from most current separation techniques. In practice, Edman sequencing is possible only after 1-D or 2-D gel electrophoresis, and then only for the most abundant protein species found.

Today, most large-scale protein identification procedures use mass spectrometry (MS) data as a starting point rather than Edman degradation. Mass spectrometry accurately determines the molecular mass of the analyzed protein. Additional information can be obtained by cleavage of the protein into smaller peptides before performing the mass spectrometry. Cleavage of proteins is usually done by enzymatic means, most commonly by trypsin which cleaves specifically the C-terminal side of arginine or lysine.

There are several identification methods from mass spectrometry data (Gras and Muller, 2001). The most widely used method consists in measuring masses of peptides resulting from the digestion process by mass spectrometry. The resulting MS spectrum represents a peptide mass fingerprint (PMF), which is characteristic for each protein. Identification by peptide mass fingerprint requires a pre-existing protein database, either directly produced or derived from a nucleic database. Identification is done by comparing the experimental masses/spectra obtained by MS (PMF) and the theoretical masses/spectra of virtually digested protein sequences present in the database. The shared masses between the experimental and theoretical spectra are used in a more or less elaborated scoring function to identify the protein. Some tools only count the number of matches, such as PepSea (Mann et al., 1993), PeptideSearch (Mann and Wilm, 1994), PeptIdent/MultIdent (Wilkins et al., 1999a; Wilkins et al., 1999b), while others use a probabilistic and/or statistic approach, such as MassSearch (Gonnet, 1992), MOWSE (Pappin et al., 1993), MS-Fit (Clauser et al., 1995), Mascot (Perkins et al., 1999), ProFound (Zhang and Chait, 2000).

Finally, the algorithm developed by Gras, SmartIdent (Gras et al., 1999; Gras et al., 2000), uses a machine learning approach.

Unfortunately, the PMF method may not always succeed in giving a reliable identification, for example when the concentration of the protein of interest is low, when only a few peptides are found after the digestion process or when the protein of interest is insufficiently purified. In addition, post-translational modifications (PTMs) or polymorphisms may modify the peptide masses and impair proper matching. Finally, it is possible that the protein of interest is simply not present in the protein database, and therefore cannot be matched.

In cases where identification is uncertain, one can use tandem mass spectrometry (MS/MS). MS/MS spectra are obtained after selection of a peptide coming from the digestion process of the protein of interest, subsequent fragmentation of said peptide (for example, by collision with a rare gas), and measurement of the produced fragment masses. Ideally, fragmentation occurs between every amino acid of the peptide, and the masses of two adjacent ionic peaks differ by the mass of one amino acid. In addition to a PMF similar to the one obtained from MS identification, MS/MS data provide information concerning the peptide sequence and allow a more detailed interpretation level than MS spectra alone.

Exploiting the information contained in MS/MS spectra is difficult due to various factors. Notably, the fragmentation process is hardly foreseeable and depends, among other things, on the amount of energy

used by the mass spectrometer, on the number and the repartition of the charges carried by the ionic fragment, on its sequence, etc..

Two main identification strategies have been devised to exploit MS/MS data: *de novo* sequencing followed by sequence matching, and direct spectrum matching with theoretical spectra from an existing database.

De novo sequencing consists in deriving a peptide sequence from its MS/MS spectrum without use of any information extracted from a pre-existing protein or nucleic database. To do so, *de novo* sequencing uses not only the mass values represented by peaks in the mass spectra, but also their position respective to each other. Early methods required generating all possible sequences whose masses are similar to the spectrum's parent mass and all the corresponding virtual spectra, PAAS3 (Sakurai et al., 1984). The experimental spectrum was then compared and matched with the virtual spectra. This approach was rapidly abandoned due to the combinatorial explosion it implies. Another strategy was to make successive possible extension of sequences (Ishikawa and Niwa, 1986). The sequences are built by successive extension with one or more amino acids. For each iteration, the sub-sequences and the corresponding virtual spectra are compared with the experimental spectrum, and the most divergent sequences are eliminated. Still another, more sophisticated strategy uses the information lying in the succession of the peaks to make the sequence extensions (Siegel and Bauman, 1988), SEQPEP (Johnson and Biemann, 1989). In this approach, the peptide sequence is built step by step, from the masses differences of "neighbor" peaks in the spectrum. This method can be viewed as the

precursor of methods based on graph representation (Bartels, 1990), (Hines et al., 1992), SeqMS (Fernandez-de-Cossio et al., 1995; Fernandez-de-Cossio et al., 1998; Fernandez-de-Cossio et al., 2000), Lutefisk97 (Taylor and Johnson, 1997; Johnson and Taylor, 2000; Taylor and Johnson, 2001), SHERENGA (Dancik et al., 1999), (Chen et al., 2001). The vertices in the graph are built from the peaks of the spectrum and represent masses of potential fragments. Physico-chemical properties are taken into account to associate a score to each vertex. Whenever two vertices differ by the mass of one or several amino acid, they are connected by an arc. Therefore, each path in the graph represent a possible sequence that can be built from the spectrum. Special algorithms then search the graph for the best paths (i.e. having the highest score built from the vertices score belonging to the path), allowing to determine the most probable sequence or sequences corresponding to the experimental spectrum. Accordingly, *de novo* sequencing results in one or a limited number of possible amino acid sequence, obtained without any recourse to a protein or nucleic database.

For identification purposes, the sequence(s) (partial or complete) obtained *de novo* are then used to scan a protein database with a standard alignment software. *De novo* sequencing is a fairly complex task which requires both good quality spectra and manual verification by a mass spectrometry expert. Accordingly, this approach is not adapted to the huge amounts of data generated by high-throughput settings available today.

The alternative to *de novo* sequencing is to match the experimental peptide spectra obtained from MS/MS with theoretical spectra derived from pre-existing protein databases. Unlike *de novo* sequencing, most MS/MS spectra matching tools use only the mass values in the MS/MS spectra - to the exclusion of their respective positions. The method most used today for MS/MS identification is the shared peak count (SPC). The ionic masses of the MS/MS spectrum represent an "ion mass fingerprint", by analogy with the "peptide mass fingerprint". The experimental MS/MS spectrum is compared with theoretical ion mass fingerprints of virtually digested and fragmented proteins in the database. Their similarity is determined by a combination of independent scores of correlations between the experimental and theoretical common masses.

Various SPC algorithms have been developed. All are based on a probabilistic score depending on the mass errors and differ mainly by their scoring function, which can be more or less sophisticated. MSTag, PepFrag (Fenyo et al., 1998), and MASCOT (Perkins et al., 1999) are examples. One algorithm - SCOPE (Bafna and Edwards, 2001) - uses both a complex probabilistic model and a dynamic programming method. Another algorithm, SEQUEST (Eng et al., 1994; Yates et al., 1995; Yates et al., 1996; Gatlin et al., 2000), uses two filtering levels: SPC followed by cross-correlation by means of fast Fourier transformation. Concerning modifications, any mutation or PTM of the source protein is susceptible to drastically modify the MS/MS spectra in comparison to the unmodified protein in the reference database: modified fragment masses are shifted by a delta corresponding to the mass difference brought by the

modification/mutation. As a result, a source modified peptide might not find any corresponding match in the reference protein database. SPC methods generally include in the database all modified/mutated peptides that they want to consider, which requires prior knowledge of the mass difference associated with the modifications/mutations taken into account. Accordingly, modifications whose mass difference with the unmodified peptide is unpredictable (such as glycosylations) cannot be taken into account by SPC methods. In addition, including all possible modifications/mutations of the peptides in the database is unrealistic due to the combinatorial explosion it implies. As a result, SPC methods usually take into account only a few very common modifications occurring on specific amino acids, such as methionine oxidation or cysteine carbamidomethylation.

In addition to the combinatorial problem, SPC algorithms have two other limitations. First, they consider the peaks independently of each other, thereby losing some important information contained in MS/MS spectra. Second, SPC algorithms need to allow a large error tolerance when used with badly calibrated spectra. As a result, the high intrinsic accuracy of current mass spectrometers is basically lost.

Two non-SPC methods have been described: spectral convolution and spectral alignment, with PEDANTA (Pevzner et al., 2000; Pevzner et al., 2001) their corresponding tool, which are claimed to be very efficient in dealing with modifications/mutations, including unpredictable modifications. Indeed, they have a major advantage over SPC methods, because they use logical constraints imposed by the spectrum peak

composition to limit the number of considered modifications/mutations. One obvious trade-off of these approaches is that one must parse the whole peptide database without using the parent mass as filtering. In addition, the combinatorial problem grows with the number of contemplated mass shifts. Accordingly, the number of modifications/mutations considered must be kept sufficiently low in order to allow identifications that are sufficiently discriminating.

SUMMARY OF THE INVENTION

According to the present invention, tandem spectrometry data (MS/MS data) obtained experimentally from peptide and/or protein-containing samples is interpreted and structured in a way allowing full exploitation of the information contained in it during matching of the structured data with biological sequence database.

DESCRIPTION OF THE DRAWING

Fig. 1 is a flow chart showing the general pathway of the method for identifying peptides or proteins from MS/MS data according to an embodiment of the present invention.

DESCRIPTION OF THE INVENTION

The present invention concerns a peptide and protein identification method using MS/MS data, obtained by any standard or non-standard

method of tandem spectrometry, such as, for example, ESI/MALDI Q-TOF MS, ESI/MALDI Ion-Trap MS, ESI triple quadrupole MS or MALDI TOF-TOF MS. Instead of directly comparing the experimental MS/MS spectrum with theoretical sequences from the database as in SPC, the method of the present invention compares an interpreted and structured view of the experimental MS/MS spectrum with theoretical sequences.

In the method of the invention and referring to Figure 1, one first performs tandem spectrometry on a sample 0, containing one or more protein or peptide. The MS/MS spectrum is then translated into a peak list 1, listing discrete mass peaks. This step can be performed by standard mass spectrometry equipment. The resulting peak list 1 is then interpreted into a list of possible mass explanations (interpreted peak list 2) taking into account physico-chemical knowledge, notably concerning the mass spectrometer, fragmentation energy levels and chemical notions (ion type, charge number, etc.). The interpreted peak list 2 is then transformed into a structured representation 3, taking into account biological knowledge - notably amino acid properties -, and preserving at least the following information:

- Mass/charge ratio of the peaks
- Mass/charge ratio of the parent peptide
- Charge of the parent peptide
- Intensity of the peaks

Identification of the peptide is performed by matching said structured representation with a biological sequence database. Said database 4 is built from any source of biological sequences 5 such as a nucleic

database translated into a protein or peptide database, or any subset of such databases. A number of sequence libraries can be used, including for example GenBank (Benson et al., 2002), EMBL (Stoesser et al., 2002), DDBJ (Tateno et al., 2002), SWISSPROT (Bairoch and Apweiler, 2000), and PIR (Barker et al., 2000). The matching with the biological sequence database is performed prior to any reduction of the structured representation 3 into one or a limited number of amino acid sequences, in contrast to *de novo* sequencing. The matching process leads to a similarity score 8 for each peptide sequence. This score is then used to determine the best peptide match or matches 9.

The present invention also provides a protein identification method comprising the steps of the peptide identification method just described, and comprising a further step consisting in using the peptide matching information for identification of the corresponding protein or proteins in a protein database.

In a preferred embodiment of the invention, the structured representation matched with the database is a graph 3 wherein vertices 6 of the graph 3 represent "ideal" fragments, built from MS/MS peaks (in the interpreted peak list 2) under a ionic hypothesis. Each vertex 6 representing a fragment indicates among others the molecular mass value of said fragment, the specific ionic hypothesis (ion type) for this fragment, and is assigned a score value expressing the credibility level for the vertex. Two vertices 6 are connected by an edge 7 whenever their mass difference is equivalent to the mass value of one or more amino acids, depending on the combinatorial level chosen.

Letters representing these specific amino acids are attached to the edge 7. Accordingly, the graph 3 represents all amino acid tags and complete sequences that can possibly be built from the MS/MS spectrum. Identification of the best peptide match or matches 9 is performed using the similarity scores 8 obtained by comparing theoretical peptides from the peptide sequence database 4 and the graph 3.

The method of the present invention compares the structured representation (or graph) 3 with theoretical peptides from a peptide sequence database 4. In contrast to identification by *de novo* sequencing followed by sequence matching - that uses database information only after reduction of the graph to one or several sequences -, the present invention directly uses database information to direct the comparison with the structured representation or graph. The goal is to find sections (sets of consecutive edges 7) of the structured representation or graph 3 which best explain the peptide. Although a section can be viewed as a classical tag encompassing sequence information, it is more than that as it contains additional information used in the comparison process.

In the present invention, the structured representation in general, and the graph structure in particular, have significant advantages over existing methods. This approach first eliminates the calibration issue during the comparison process. As already mentioned, peak masses in MS/MS spectra can be shifted of a significant value in spite of the high intrinsic accuracy of the spectrometer. As a result, existing identification methods based on SPC must allow for a high tolerance

error when comparing peak masses and theoretical fragment masses, which leads to a significant increase of the noise level, hence of the number of false positives. The method of the present invention compares differences of peak masses with differences of theoretical masses. Because differences of adjacent masses are weakly influenced by calibration errors, the method of the present invention allows to fully take advantage of the spectrometer accuracy. Another advantage of the structured representation is that it allows to take into account not only the number of peak matches (as in SPC), but also the number of successive matches susceptible to explain the sequence.

In a preferred embodiment of the invention, the matching of the structured representation with sequences in the database is performed by parsing the structured representation or the graph according to each database sequence, each parsing leading to a score correlating each database sequence to the structured representation or graph.

This approach allows notably to compare the structured representation with any sub-sequences of the peptide sequence database, each parsing leading to a score correlating the sub-sequence with a section of the structured representation or graph. In case of incomplete spectral information, non-linked relevant sets of successive edges (sections) can be combined together to form a same peptide sequence. In case of modified source peptides, this approach also allows to combine non-linked relevant sets of successive edges (sections) according to a modification hypothesis.

Representations under a graph structure allow to keep all the original information, as well as to consider information coming from many different sources during the comparison process. The graph includes two information types : first, local information, which are used for the path building in order to favor most pertinent edges and which are stored in variables associated with vertices and edges (as the vertices mass, intensity, score or the edge amino acid), and second, global information, which describe path pertinence related to the current peptide or to any subsequence belonging to it, and possibly stored in weights associated with edges. Local and global parameters must be weighted and combined in a way maximizing the performance of the identification algorithm, and allowing sufficient discrimination between the peptide ranked first and the other candidates. Using a set of identified spectra from a known mass spectrometer, it is possible to optimize the weights with genetic algorithms (Gras et al., 2000; Gras et al., 1999).

In another embodiment of the invention, said parsing is performed through the use of a Swarm Intelligence-type algorithm (Kennedy and Eberhart, 2001; Bonabeau et al., 1999). Swarm intelligence is a form of distributed artificial intelligence: self-organization of unsophisticated units - agents -, evolving and interacting within a given environment and able to manage direct and/or indirect communication, results in the emergence of an intelligent collective behavior.

In still another embodiment of the invention, the Swarm Intelligence-type algorithm is an algorithm called "Ant Colony Optimization" (ACO) (Dorigo and Di Caro, 1999). ACO algorithms are defined as multi-agent systems inspired from real ant colony behavior. The principle of ACO is to explore, iteratively and simultaneously, different solutions of a given problem by an ant-agent population. The emergent collective behavior is guided by indirect communication between the ants, mediated by environmental modifications (stigmergy). Ants modify their environment by depositing given amounts of pheromone, which are locally accessible and affects the behavior of the other ants. In this embodiment, an ACO algorithm inspired from the "trail-laying/trail-following" foraging behavior of ants is used to score the matching of current peptide of the database with the structured representation. Since ants can find the shortest path connecting the colony to the food source, it is possible to exploit the rules governing the foraging process and use them to find good scoring paths in the graph. Each ant obtains a score depending on the quality of the found solution. The use of virtual pheromone allows good solutions to be memorized and act as a positive feedback (intensification of the search). In order to avoid premature convergence, a certain amount of pheromone also evaporates at each iteration (negative feedback, diversification of the search).

The modified ACO used to parse the graph first sets the pheromone quantity of each edge to a tiny value. Then, the ants parse the graph iteratively. At each iteration, the ants move on the graph from one vertex to the other, using existing edges or, if allowed, jumping from one vertex to the other until a stop criterion is reached (for example, when arrived on a vertex having no successor). The choice of the next

edge results from a probabilistic computation, taking into account both local parameters (i.e. the score of the successor vertex) and the global learning already done (i.e. the amount of pheromone on the successor edge). At the end of each iteration, some pheromone is automatically removed from each edge (evaporation), while some pheromone is added on each edge parsed by an ant (the exact amount being dependent on the ant's score). As a result, the algorithm allows gradual convergence toward one or several good scoring sections, which can be further correlated in order to maximally cover the theoretical candidate peptide, ultimately leading after analysis of all peptides to a ranked list of candidate peptides.

The ACO algorithm has several advantages. For example, the stochastic nature of the ant motion allows to parse any path in the graph. All possible mutations compatible with the MS/MS spectrum are implicitly represented in the graph, and possible modifications can be contemplated by allowing the ants to jump from one vertex to another, unconnected one. Like spectral alignment methods, the present invention uses the spectrum logical constraints to limit the combination number of possible modifications. In addition, it drastically restricts this number by allowing only directed jumps joining relevant sections of the representation or graph. Thus, only modifications enhancing the global correspondence between the sequence and the spectrum are considered. It is also possible to restrict the vertices allowed for an ant, depending on the vertices already parsed by this ant. This allows to accept, for example, only one missed-cleavage : an ant having used an edge corresponding to a lysine could avoid to further incorporate a second lysine.

An additional advantage of the present invention is that switching from it to a more traditional *de novo* sequencing mode is straightforward, by simply letting aside the information coming from the database.

The invention also provides a system comprising a computer linked to one or more mass spectrometers and one or more biological sequence databases, said computer comprising a program for performing the steps of the methods described herein.

The invention also provides a computer-readable medium comprising instructions for causing a computer linked to one or several mass spectrometers and to one or more biological sequence databases to perform the steps of the methods described herein.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT

The following paragraphs provide a detailed description of MS/MS data treatment and identification according to a preferred embodiment of the invention, combining a graph representation and an ACO algorithm and called Popitam (Peptide Or Protein Identification from TAndem Mass spectrometry).

I. Peak interpretation

Let us define $S_{exp}=\{s_1, s_2, \ldots, s_{|sexp|}\}$, the experimental MS/MS peak list to identify, and a set of ionic hypothesis $\Delta=\{\eta_1, \eta_2, \ldots, \eta_{|\Delta|}\}$. A ionic hypothesis can be seen as a possible interpretation of a peak. Each η_i has four attributes, which are presumptions concerning the ionic fragment s_j, measured by the spectrometer : an offset value $o(\eta_k)$, i.e. the mass difference between the ionic fragments and the corresponding b-ion type fragment (for comprehension purpose, we will call such fragments *b-fragments*, and their corresponding masses *b-masses*), a terminus side $t(\eta_k)$ (N-term or C-term), a number of charges $c(\eta_k)$, and an approximated occurrence probability $p(\eta_k)$. The probability $p(\eta_k)$ depends among other things on the spectrometer used, and can be determined during a learning phase using a set of identified spectra (Dancik et al., 1999).

The interpretation process consists in attributing to each peak from S_{exp} a ionic hypothesis comprising all four attributes described above. Therefore, each peak s_j from S_{int} will be characterized by a mass/charge ratio $\mu(s_j)$, an intensity $\iota(s_j)$, and a ionic hypothesis $\eta(s_j)$. The number of elements in the interpreted peak list S_{int} is : $|S_{int}| = |S_{exp}| \cdot |\Delta|$. This approach means that at least $|\Delta| - 1$ interpreted peaks computed from a given peak in S_{exp} are false.

II. Graph construction

Let us define a spectrum graph $G=(V,E)$ as a directed acyclic graph, with a set of vertices $V=\{v_1, v_2, \ldots, v_{|V|}\}$ and of edges $E = \{e_{ij} | i<j<|V|, v_i \text{ and } v_j \in V\}$. Each vertex v_i is characterized by a b-mass, $\mu(v_i)$ and its corresponding ionic peak mass/charge ratio $\mu^s(v_i)$, an intensity $\iota^s(v_i)$, a score $\sigma(v_i)$, a ionic hypothesis $\eta(v_i)$, a family $F(v_i)$, and a successor list $succ(v_i)$, while each edge $e_{ij} \in E$ is characterized by a pheromone trail $\tau(e_{ij})$ and a label $\lambda(e_{ij})$.

II. 1) Building the vertices :

G is built from the peak list S_{int}. The first step is to transform all interpreted peaks into b-ions charged once, which represent N-terminal "ideal" fragments.

Each peak from S_{int} leads to a vertex v_i. Given M_{exp} the experimental parent mass, with $M_{exp}=(M_{obs}-1)\cdot c(M_{obs})$, M_{obs} being the mass/charge ratio of the peptide parent mass, and $c(M_{obs})$ its charge number, we built the vertices according to algorithm 1.

Algorithm 1 : Building the vertices

```
i = 0;
For each s_j ∈ S_int {
    if (t(η(s_j)) == "N-term")
        μ(v_i) ← c(η(s_j))·μ(s_j) − (c(η(s_j)) − 1) − o(η(s_j));
    if (t(η(s_j)) == "C-term")
        μ(v_i) ← M_exp − [c(η(s_j))·μ(s_j) − (c(η(s_j)) − 1) − o(η(s_j))];
    μ^s(v_i) ← μ(s_j);
    ι^s(v_i) ← normalize(ι(s_j));
    i++;
}
```

We also create an initial vertex corresponding to the empty sequence and a final vertex corresponding to the complete sequence. Therefore, the number of vertices is equal to $|S_{int}|+2$.

II. 2) Vertex families :

For each vertex, a family F of neighbor vertices is defined. The concept of family is based on the idea that when a b-fragment is represented by several ionic peaks in S_{exp}, the computed b-masses $\mu(v_i)$ of theses peaks will be almost equal. The family building is hence based on the vertex b-mass differences, which must be lower than a specified threshold. We chose not to merge the vertices as described in (Dancik et al., 1999), because the merging process does not manage the calibration error on the peaks and depends on the parent mass accuracy, which is often quite low. Accordingly, two b-masses representing the same b-fragment and derived by ionic hypothesis of different terminal types ($t(\eta(v_i))\neq t(\eta(v_j))$) can be quite different when compared to the b-masses obtained from ionic hypothesis of same terminal type. Such b-masses therefore cannot be merged because there are too different or, if merged can produce a new vertex with a substantially less accurate b-mass. In order to avoid this problem we do not merge the vertices, but build vertex families $F(v_i)=\{v_j \ldots v_{|F(v_i)|}\}$ containing all neighbor

vertices possibly belonging to the same b-fragment. This approach allows to keep the b-mass of the vertices unchanged, and hereby fully benefit of the accuracy of the spectrometer. In addition, the algorithm used for building the families is not greedy - as is the merging algorithm proposed by Dancik -, but is exact.

A vertex v_j is added to a family $F(v_i)$ according to the following rules. First, the two vertex b-masses must be close enough. As shown in equation 1, the threshold must be adapted, depending on whether the two vertices joined in a same family are derived by ionic hypothesis of a same terminal type or of different terminal types.

Equation 1 : $\left| \mu(v_j) - \mu(v_i) \right| < \varepsilon$

with $\varepsilon = \varepsilon_1$ if $t(\eta(v_i)) = t(\eta(v_j))$, $\varepsilon = \varepsilon_2$ if $t(\eta(v_i)) \neq t(\eta(v_j))$ and $\varepsilon_1 < \varepsilon_2$

Second, the two vertex b-masses have to be issued from different ionic hypothesis ($\eta(v_i)$!= $\eta(v_j)$).

Algorithm 2 : Building the families

```
For i = 1 to |V|
      F(v_i) = Ø;
      test1 = TRUE;
      while (test1 ) {
            v_j <- find the new closest vertex(v_i);

            if (term(v_i) == term(v_j))      ε = ε_1;
            else                             ε = ε_2;

            if (|v_j - v_i| < ε) {
                  test2 = TRUE;
                  For each v_k ∈ F(v_i)
                        if (η(v_k) == η(v_j)) : test2 = FALSE;

                  if (test2) : F(v_i) = F(v_i) ∪ v_j;

            }
            else test1 = FALSE;
      }
}
```

II. 3) Scoring the vertices :

Because the vertices are built under some assumptions, we need a value defining the credibility level of each vertex. This value is represented by a score $\sigma(v_i)$, defined according to a non exhaustive list of criterions. Two criterions are currently taken into account, leading to a redundancy score $\rho(v_i)$ and a probability score $\pi(v_i)$.

Equation 2 : $$\sigma(v_i) = \sqrt[\rho(v_i)]{\pi(v_i)}$$

Once the families are defined, it is possible to compute $\rho(v_i)$ and $\pi(v_i)$. The redundancy score $\rho(v_i)$ must be increased according to the family size as several equivalent b-masses confirm the ionic hypothesis of v_i, while the probability score $\pi(v_i)$ takes into account the occurrence probability $p(\eta)$ of the family members :

Equation 3 : $$\pi(v_i) = \prod_{v_j \in F(v_i)} p(\eta(v_i)) \cdot \prod_{v_j \notin F(v_i)} (1 - p(\eta(v_i)))$$

II. 4) Connecting the graph :

If the b-masses of two associated vertices v_i and v_j differ by the value of one or several amino acids, they can be connected by an edge e_{ij}. According to the number of amino-acids included in a given edge,

the latter can be called a simple edge ($|\lambda(e_{ij})|=1$), a double edge ($|\lambda(e_{ij})|=2$), and so on. Let $A=\{a_1,a_2,\ldots,a_{|A|}\}$ be the alphabet of the amino-acids. A contains all common amino-acids, as well as some modified amino acids, such as carboxymethylated cysteine, carbamidomethylated cysteine, or oxidated methionine. Each $a_i \in A$ has a mass $\mu(a_i)$ and a label $\lambda(a_i)$. $A^c=\{a_1^c,a_2^c,\ldots,a_{|A^c|}^c\}$ is the set of all combinations of 1 to N amino acids among $|A|$. Because the edge number increases exponentially with the value of N, the latter is usually small (typically N=2 or N=3).

Given $\mu(a_n^c)$, the sum of the masses of all amino acids in a_n^c, and $\lambda(a_n^c)$, formed from the labels of the amino acids in a_n^c, the algorithm 3 shows the computation of the edges. The vertex list must be sorted according to the b-masses values.

Algorithm 3 : Connecting the graph

```
For  i = 0 to |V|
      For  j = i + 1 to |V| {
            if  (t(η(v_i)) == t(η(v_j)))    ε = ε_1;
            else                            ε = ε_2;
            For  n = 1 to |A^c| {
                  if  (|μ(v_j) - μ(v_i) - μ(a_n^c)| < ε)
                        createEdge   (e_ij, a_n^c); :
            }
      }
```

III. Identification process

III. 1) The peptide database

Let $D=\{P_1, P_2, \ldots P_{|D|}\}$ be the peptide database used for the identification. The peptides P_c can be obtained from the whole or a subset of nucleic or protein databases. P_c are characterized by three attributes. First, their sequence $Q(P_c)=\{a_1^P, a_2^P, \ldots, a_{|Q(P_c)|}^P\}$ with $a_n^P \in A$. Second, their theoretical mass $\mu(P_c)$ (see equation 4). Third, an identification score $score(P_c)$.

Given the terminus mass values μ(N-term) and μ(C-term), $\mu(P_c)$ is obtained as follows :

Equation 4 : $$\mu(P_c)=\mu(N-term)+\mu(C-term)+\sum_{n=1}^{|Q(P_c)|}\mu(a_n^P)$$

The identification process consists in comparing the peptides of D with the graph G and in correlating each peptide $P_c \in D$ with a score $score(P_c)$. Given M_{exp}, the experimental parent mass of the spectrum, and r, a predetermined threshold, we have :

Algorithm 4 : Identification process

```
For c = 1 to |D|
    If ( |μ(P_c) - M_exp| < r )
        score(P_c) = compare(P_c,G)
```

This algorithm results in a list of candidate peptides ranked by score. The following paragraph describes the *compare* function, which performs the comparing of a theoretical peptide with the graph.

820

III. 2) Comparison process

The comparison process between the graph G and a peptide P_c requires to find in G the sections best explaining P_c. A complete section is a path
825 in the graph corresponding to a whole peptide sequence. We present here a possible non deterministic strategy to search, for a given P_c, the best complete section in G. The algorithm will be modified further in order to extract sections instead of complete paths.

830 Let $F=\{f_1, f_2, \ldots, f_{|F|}\}$ be the ant population. Each ant f_k, walking on the graph at iteration t, builds a path which includes a set of vertices $L_V^t(f_k)$, subset of V, such that

$$L_V^t(f_k)=\left\{v_1, v_2, \ldots, v_{\left|L_V^t(f_k)\right|}\right\}$$

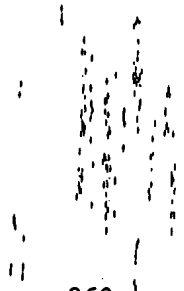
and consequently, a set of edges, denoted $L_E^t(f_k) \subset E$ of size $\left|L_E^t(f_k)\right|$. The
835 quality of $L_E^t(f_k)$ is represented by the ant's score $S^t(f_k)$. The concatenation of the edge labels $\lambda(e_{ij})$, with $e_{ij} \in L_E^t(f_k)$, represents the sequence

$$L_Q^t(f_k)=\left\{a_1^L, a_2^L, \ldots, a_{\left|L_Q^t(f_k)\right|}^L\right\} \quad , \quad a_i^L \in A^c$$

built by ant k.

840

Algorithm 5 is an adaptation to our problem of an ACO algorithm. First, $\tau(e_{ij})$, the amount of pheromone of each edge $e_{ij} \in G$ is initialized (with $\tau_0=10^{-6}$), as well as the best complete path found in the graph (L^*) and its associated score $S(L^*)$. At the beginning of each iteration (t_{max} is the predefined total number of iterations), the amount of pheromone that will be added at each edge, $\Delta\tau(e_{ij})$, is initialized at 0. Then, each ant parses the graph, building its own path $L_E^t(f_k)$ and gets a score $S^t(f_k)$. This score is used for updating the $\Delta\tau(e_{ij})$ for each $e_{ij} \in L_E^t(f_k)$. Q is a predefined constant value, chosen of a same order of magnitude as that of the optimal score. Authors have demonstrated that the value of Q has little influence on the final result (Theiler, 2001; Bonabeau et al., 1999). If the path built by the ant obtains a higher score than $S(L^*)$, L^* and $S(L^*)$ are updated. Finally, when all ants have parsed the graph and have added their contribution to the $\Delta\tau(e_{ij})$, the graph is updated, $\omega \in [0;1[$ being the evaporation rate. At the end, the *compare* function returns the score of the best path attributed to P_c.



Algorithm 5 : Finding the best path in G for a peptide P

Initiation :

$L^+ = \emptyset$;

$S(L^+) = 0$;

For each edge $e_{ij} \in E : \tau(e_{ij}) = \tau_0$

Iterations :

For $t = 1$ to t_{max} {

For each $e_{ij} \in E : \Delta\tau(e_{ij}) = 0$;

For $k = 1$ to $|F|$ {

$\left(L_V^t(f_k), L_E^t(f_k), L_Q^t(f_k)\right) = \text{parseGraph}(P_c, f_k)$;

$S^t(f_k) \leftarrow \text{scoreAnt}\left(P_c, f_k, L_V^t(f_k), L_E^t(f_k), L_Q^t(f_k)\right)$;

For each $e_{ij} \in L_E^t(f_k) : \Delta\tau(e_{ij}) = \Delta\tau(e_{ij}) + \frac{S^t(f_k)}{Q}$; //update $\Delta\tau(e_{ij})$

if $\left(S(L^+) < S^t(f_k)\right)$ { // update best path

$S(L^+) \leftarrow S^t(f_k)$;

$L^+ \leftarrow L_E^t(f_k)$;

}

}

For each $e_{ij} \in E : \tau(e_{ij}) \leftarrow (1-\omega)\cdot\tau(e_{ij}) + \Delta\tau(e_{ij})$; // update graph

}

return $S(L^+)$;

A more detailed description of the *parseGraph* and *scoreAnt* functions follows:

III. 2a) Parsing the graph :

The ant f_k is first placed on the initial vertex v_1. It can go forward as long as the current vertex v_i has any successors ($succ(v_i) \neq \emptyset$), and as long as the length of its built sequence $|L_Q(f_k)|$ is smaller than the length of the current database sequence $|Q(P_c)|$. The transition rule used to go from a vertex v_i to a vertex v_j with $v_j \in succ(v_i)$ depends on three pieces of information. The first one is visibility, represented by $\sigma(v_j)$, the score of the successor vertex. It can be considered as a local parameter. The second piece of information corresponds to the memory of the learning previously done by the ant population. It is a global parameter, representing the amount of pheromone laid on the edge e_{ij}, $\tau(e_{ij})$. Finally, the third piece of information is the sequence of the current database peptide P_c. Indeed, if the label of the next edge e_{ij} matches the next amino acid in the sequence $Q(P_c)$, the transition probability is multiplied by a predefined constant value dependent upon the edge label length.

Given α and β, two adjustable parameters controlling the relative weight of the learning and the visibility, $p_t^{f_k}(e_{ij})$, the probability for ant f_k to take the edge e_{ij} at iteration t, $p_t^{f_k}(e_i)$ the set of these probabilities for all $succ(v_i)$, and $Q(P_c)=\{a_1^p, a_2^p, \ldots, a_{|Q(P_c)|}^p\}$, the current peptide sequence :

Algorithm 6 : Parsing G with ant f_k

$i = 1;$

$L_E^t(f_k) = \emptyset;\ L_V^t(f_k) = \emptyset;\ L_Q^t(f_k) = \emptyset;$

while $(succ(v_i) = \emptyset)$ and $\left(\left| L_Q^t(f_k) \right| < \left| Q(P_c) \right| \right)$ {

for each $v_j \in succ(v_i)$ {

$$p_t^{f_k}(e_{ij}) = \frac{\tau(e_{ij})^\alpha \cdot \sigma(e_{ij})^\beta}{\sum_{\forall succ(v_i)} \left(\tau(e_{ij})^\alpha \cdot \sigma(e_{ij})^\beta \right)};$$

if $\left(\text{match} \left(a^P_{\left|L_Q^t(f_k)\right|+1}, \ldots, a^P_{\left|L_Q^t(f_k)\right|+\left|\lambda(e_{ij})\right|}, \lambda(e_{ij}) \right) \right) : p_t^{f_k}(e_{ij}) = p_t^{f_k}(e_{ij}) \cdot c_{\left|\lambda(e_{ij})\right|};$

// here, we compare all permutations in $\lambda(e_{ij})$

with the amino acids $a^P_{\left|L_Q^t(f_k)\right|+1}, \ldots, a^P_{\left|L_Q^t(f_k)\right|+\left|\lambda(e_{ij})\right|}$

$add(p_t^{f_k}(e_i), p_t^{f_k}(e_{ij}));$

}

$normalize(p_t^{f_k}(e_i));$

$e_{ij} = chooseEdge(p_t^{f_k}(e_i));$

$add(L_V^t(f_k), v_j);$

$add(L_E^t(f_k), e_{ij});$

$add(L_Q^t(f_k), \lambda(e_{ij}));$

$i \leftarrow j;$

}

III. 2b) Scoring the ants

At the end of each iteration t, one must evaluate the similarity between the current peptide P_c and the different paths used by the ants. Each ant gets a final score $S^t(f_k)$ depending on its path $L_E^t(f_k)$. The goal is to include in $S^t(f_k)$ all possibly relevant information from different sources (see equation 5). For example, in order to take into account information coming from S_{int} we can use the intensity of the

peaks, stored in $\iota^s(v_i)$, $v_i \in L_v^t(f_k)$, and compute an intensity score intS. From the ionic hypothesis set, we can build a relevancy score relS, expressing the relevancy of the vertices parsed by f_k. The current peptide sequence can be used in a covS score that would express the similarity between the peptide sequence $Q(P_c)$ and the sequence $L_Q^t(f_k)$ built by the ant. The quality of the correlation between the b-masses of the used vertices and the theoretical masses expected from $Q(P_c)$ can also be taken into account as a regression score called regS. Still other information can be added, such as rules resulting from the expertise of biologists used to studying MS/MS data.

Equation 5 : $S^t(f_k) = f(intS, relS, covS, regS, ...);$

The next sections show implementation examples of the sub-scores intS, relS, covS and regS used in our current algorithm.

The coverage score recS represents the sequence similarity between the current peptide P_c and the sequence built by an ant f_k. It is computed with an alignment function as for example a Smith and Waterman algorithm. Given $Q(P_c)$ and $L_Q^t(f_k)$:

Algorithm 7 : Coverage score

$$recS = align\left(Q(P_c), L_Q^t(f_k)\right);$$

The relevancy score is the mean of the used vertices score. It is computed as shown in equation 6.

Equation 6 : $$relS = \frac{\sum_{v_i \in L_V^t(f_k)} \sigma(v_i)}{\left| L_V^t(f_k) \right|}$$

Similarly, the intensity score is computed as follows:

Equation 7 : $$intS = \frac{\sum_{v_i \in L_V^t(f_k)} \iota^s(v_i)}{\left| L_V^t(f_k) \right|}$$

The regression score measures the global correspondence between the experimental masses $\mu^s(v_i)$ of the vertices included in the ant's path and the corresponding theoretical masses $R(P_c)=\{r_1, r_2, \ldots, r_{|R(Pc)|}\}$ computed from the current database peptide sequence $Q(P_c)$ (Gras et al., 2000). The relation between these masses is first plotted on a graph, with the experimental masses as abscissa and the theoretical masses as ordinate, and the set of points allows to calculate a linear regression. The mean of the deviation between the points and the linear regression represents the regression score regS .

Given y = ax+b, the linear regression, $\mu^s(v_i) \in L_V^t(f_k)$ the experimental masses and their corresponding theoretical masses $r_i \in R(P_c)$:

Algorithm 8 : Computation of regS

For each $\mu^s(v_i) \in L_V^t(f_k)$ {

add(R, $\mu^s(v_i)$, $Q(P_c)$); // compute the corresponding theoretial mass r_i and add it to R

linearReg(a,b,R,$L_V^t(f_k)$); // this function makes the regression

$$regS = \frac{\sum_{i=0}^{|L_V^t(f_k)|} \left(a * r_i - \mu^s(v_i) + b\right)^2}{\left| L_V^t(f_k) \right|}$$

}

EXPERIMENTAL EXAMPLE

A preliminary implementation of our algorithm has been tested on a training set of MS/MS spectra (only complete paths, no unknown modifications). 92.1% of 101 spectra were well identified. Here are some result examples.

MSMS file : DSNNLXLHFNPR.dta
Peaks used/tot : 56 / 935
Parent_mass (M/H+)/charge: 1485.63 / 2
Vertices : 170
Edges (simple/double) : 482 / 4345
Ants nb / Iter nb : : 101 / 5

#	s_n*	fin_s**	access	id	sequence_dtb/sequence_graph
1.	0	1.396	P09382	LEG1_HUMAN	DSNNLCLHFNPR***
					sdNNLXLHFNPR****
2.	0	0.312	Q05586	NMZ1_HUMAN	FANYSIMNLQNR
					ewNIsinmLPNR
3.	0	0.252	P09848	LPH_HUMAN	DPSNQEDVEAARR
					rxLNQEvdaePR

* s_n = start node
** fin_s = final score
*** theoretical sequence read in the database
**** sequence parsed in the graph (uppercase = simple edge, lower case = double edge)

MSMS file : EFTNVYIK.dta
Peaks used/tot : 40 / 260
Parent_mass (M/H+)/charge: 1012.51 / 2
Vertices : 122
Edges (simple/double) : 349 / 3153
Ants nb / Iter nb : : 74 / 5

#	s_n	fin_s	access	id	sequence_dtb/sequence_graph
1.	0	1.970	Q13310	PAB4_HUMAN	EFTNVYIK
					EFTNVYIK
	0	1.970	Q15097	PAB2_HUMAN	EFTNVYIK
					EFTNVYIK
	0	1.970	P11940	PAB1_HUMAN	EFTNVYIK
					EFTNVYIK
2.	0	1.079	P42694	Y054_HUMAN	QDYEMALK
					QDeyaoLK
3.	0	0.677	P46821	MAPB_HUMAN	LKHLDFLK
					LKlhdfLK

MSMS file : EQIVPKPEEEVAQK.dta
Peaks used/tot : 64 / 317
Parent_mass (M/H+)/charge: 1622.83 / 3
Vertices : 194
Edges (simple/double) : 579 / 4566
Ants nb / Iter nb : : 120 / 5

#	s_n	fin_s	access	id	sequence_dtb/sequence_graph
1.	0	1.374	P18621	RL17_HUMAN	EQIVPKPEEEVAQK
					qeviPKPEEEVAQK
2.	0	0.396	P36383	CXA7_HUMAN	LLEEIHNHSTFVGK
					LLEEvkCHSvzVG
3.	0	0.394	P16991	YB1_HUMAN	RPENPKPQDGKETK
					RPtdPKPQvxgiQK

CLAIMS

1. A peptide identification method comprising the following steps:

(a) Performing tandem mass spectrometry on a sample containing one or more protein or peptide.

(b) Reducing the resulting spectrum to a peak list.

(c) Listing possible interpretations for said peak list into an interpreted peak list, taking into account physico-chemical knowledge.

(d) Structuring said interpreted peak list into a structured representation taking into account biological knowledge and preserving at least the following information:

- Mass/charge ratio of the peaks obtained in step (b)
- Mass/charge ratio of the parent peptide
- Charge of the parent peptide
- Intensity of the peaks

(e) Matching said structured representation with a biological sequence database prior to any reduction of the structured information into one or a limited number of amino acid sequences.

(f) Determining the best peptide match or matches within said database.

2. A protein identification method comprising steps (a) to (f) of claim 1, and further comprising a step (g) consisting in using the

peptide matching information of step (f) for identification of the corresponding protein or proteins in the protein database.

3. The method of claim 1 or 2 wherein the structured representation of step (d) consists in a graph wherein:

- Vertices of the graph represent individual elements of the interpreted peak list, translated into potential b-ion type peptide fragments.
- Edges link vertices representing said b-ion type peptide fragments whose molecular weights differ by a value equivalent to the molecular weight of one or more amino acids.

4. The method of anyone of claims 1 to 3 wherein the matching of step (e) consists in successively parsing the structured representation of step (d) according to each database sequence, each parsing leading to a score correlating each database sequence to the structured representation.

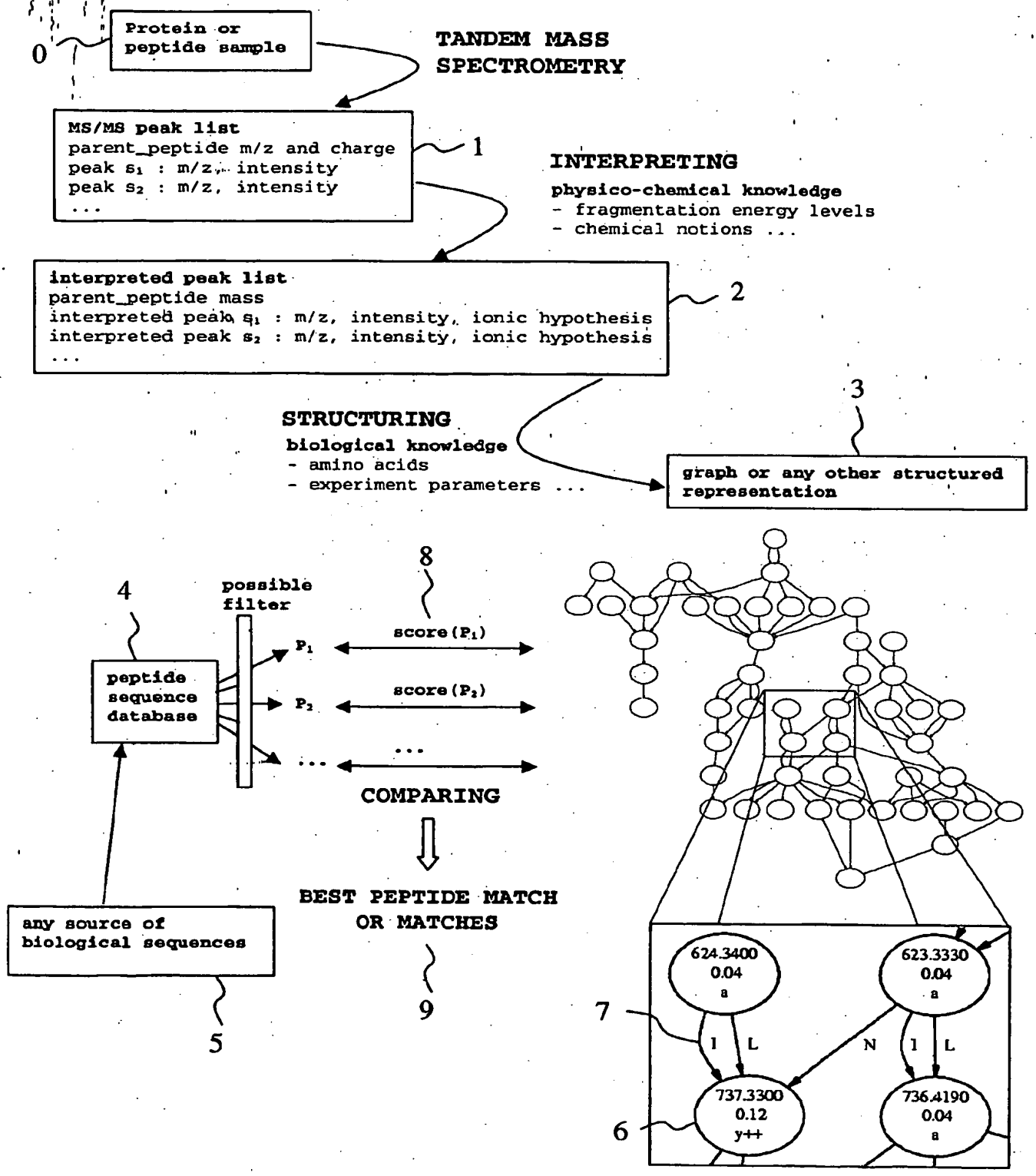
5. The method of claim 4 wherein the parsing is performed by a Swarm Intelligence Algorithm.

6. The method of claim 5 wherein the Swarm Intelligence algorithm is an Ant Colony Optimization algorithm.

7. The method of anyone of claims 3 to 6 wherein non-linked relevant sets of successive edges are combined together according to a modification hypothesis.

8. A computer-readable medium comprising instructions for causing a computer linked to one or several mass spectrometers and to one or more biological sequence databases to perform the steps of the method of anyone of claims 1 to 7.

9. A system comprising a computer linked to one or more mass spectrometers and to one or more biological sequence databases, said computer comprising a program for performing the steps of the method of anyone of claims 1 to 7.


0
Protein or peptide sample
TANDEM MASS SPECTROMETRY
MS/MS peak list
parent_peptide m/z and charge
peak s1 : m/z, intensity
peak s2 : m/z, intensity
...
1
INTERPRETING
physico-chemical knowledge
- fragmentation energy levels
- chemical notions ...
interpreted peak list
parent_peptide mass
interpreted peak s1 : m/z, intensity, ionic hypothesis
interpreted peak s2 : m/z, intensity, ionic hypothesis
...
2
3
STRUCTURING
biological knowledge
- amino acids
- experiment parameters ...
graph or any other structured representation
4
possible filter
peptide sequence database
8
score(P1)
score(P2)
...
P1
P2
...
COMPARING
BEST PEPTIDE MATCH OR MATCHES
9
any source of biological sequences
5
7
624.3400
0.04
a
623.3330
0.04
a
I
L
N
I
L
737.3300
0.12
y++
736.4190
0.04
a
6

INTERNATIONAL SEARCH REPORT

PCT/IB 02/02731

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F19/00 G01N33/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, BIOSIS, INSPEC, IBM-TDB, MEDLINE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GRAS R ET AL.: "Improving protein identification from peptide mass fingerprinting through a parametrized multi-level scoring algorithm and an optimized peak detection" ELECTROPHORESIS, vol. 20, no. 18, 1999, pages 3535-3550, XP002902845 cited in the application	1,2,8,9
Y	the whole document	3,4,7
Y	US 2002/087275 A1 (JIANG SHAN ET AL) 4 July 2002 (2002-07-04) the whole document	3,4,7
A	WO 99 62930 A (MILLENNIUM PHARM INC) 9 December 1999 (1999-12-09) the whole document	1-4,7-9

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 July 2003	22/07/2003

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Godzina, P

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

PCT/IB 02/02731

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 02 21139 A (OXFORD GLYCOSCIENCES UK LTD ;ROBINSON ANDREW WILLIAM (GB); TOWNSEN) 14 March 2002 (2002-03-14) the whole document	1,2,4,8, 9
A	BAFNA V ET AL: "SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database." BIOINFORMATICS (OXFORD, ENGLAND) ENGLAND 2001, vol. 17 Suppl 1, 2001, pages S13-S21, XP002247078 ISSN: 1367-4803 the whole document	1,2,8,9

INTERNATIONAL SEARCH REPORT

PCT/IB 02/02731

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2002087275 A1	04-07-2002	AU 7808901 A	13-02-2002
		WO 0211048 A2	07-02-2002
WO 9962930 A	09-12-1999	AU 4228499 A	20-12-1999
		WO 9962930 A2	09-12-1999
WO 0221139 A	14-03-2002	AU 8605901 A	22-03-2002
		EP 1317765 A2	11-06-2003
		WO 0221139 A2	14-03-2002
		US 2002102610 A1	01-08-2002